**Applied Multivariate Analysis**

**Prof. Amit Mitra**

**Prof. Shramishtha Mitra**

**Department of Mathematics and Statistics**

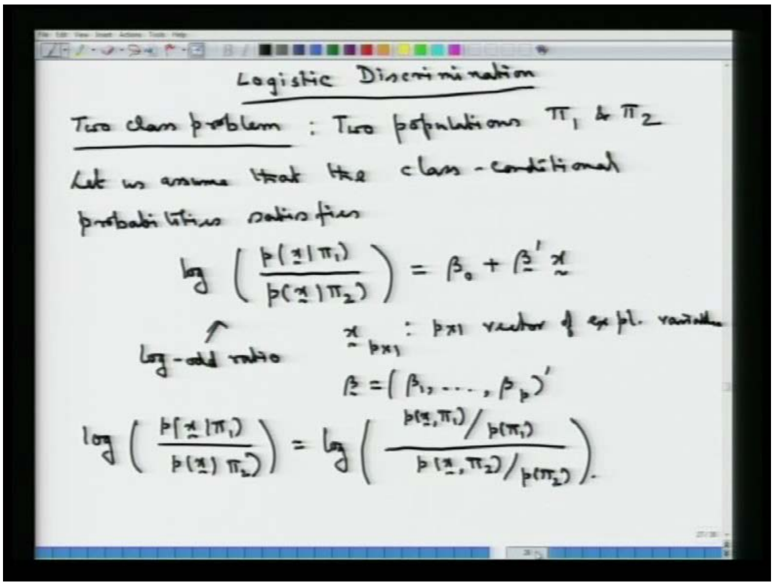
**Indian Institute of Technology, Kanpur**

**Lecture No. #35**

**Discriminant Analysis and Classification**

In this lecture, we are going to talk about logistic discrimination. So, we are going to build discrimination function that is going to be based on the principal of logistic regression. So, let us try to look at what we are up to.

(Refer Slide Time: 00:27)



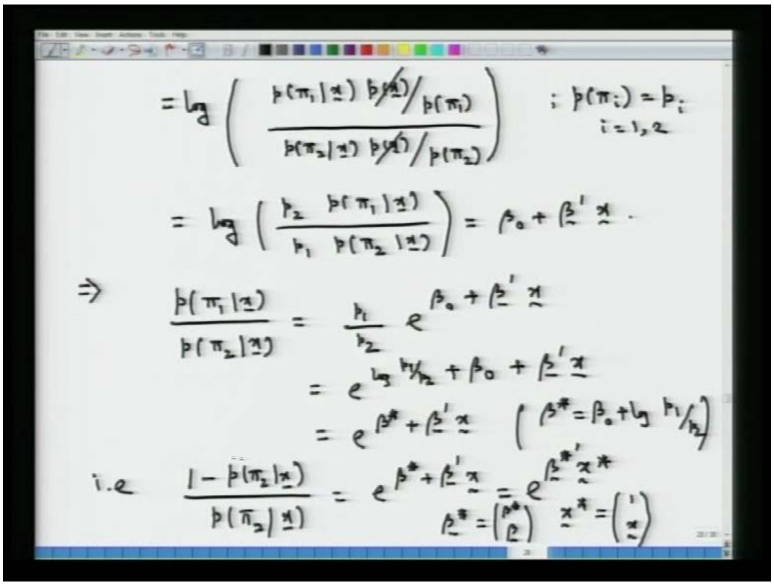

So, we are talking about construction of logistic discrimination. To start with a simple problem, let us consider a two class problem meaning thereby we have got two populations; namely pi 1, and pi 2. And we assume that the class conditional probabilities satisfy the following relationship.

Let us assume that the class conditional probabilities (No audio from 01:12 to 01.21) satisfies the following relationship that, log of p x given pi 1. So, this is the density, this is conditioned on the fact that x is coming from pi 1, that divided by the density of x, when it is coming from pi 2. So, this is the class conditional probability ratio of that this, this is the odds ratio. And then, we look at the log of that it is what is called the log odds ratio. So, we assume that, that is given by a constant beta naught plus a beta transpose vector times x.

Now, this x is what we have as say a p dimensional vector of explanatory variables p by one vector of explanatory variables, that is its basically the feature vector corresponding to which we are going to base our classification on. And this beta vector accordingly is a vector of constants beta 1, beta 2, beta p of the same order of this feature vector here, and beta naught is a constant. Now this quantity the log of this odds ratio is what is called the log odds ratio.

Now, we look at this log odds ratio and see what we get from this log odds ratio. It is probability density x given pi 1 that divided by p x given pi 2. So, if we write this log odds ratio the numerator and the denominator of these terms. Now this can be written as the joint density of x pi 1 that divided by p pi 1, which is the prior probability of the first population that divided by p x pi 2 divided by p pi 2, the prior probability of the second population.

(Refer Slide Time: 03:36)




Now, we can write this in the following way, this is further equal to log of we write it as p pi 1, the posterior density this into p x say that divided by what we had out there p pi 1, this divided by p pi 2 given x this into the marginal density of x, this divided by p pi 2.

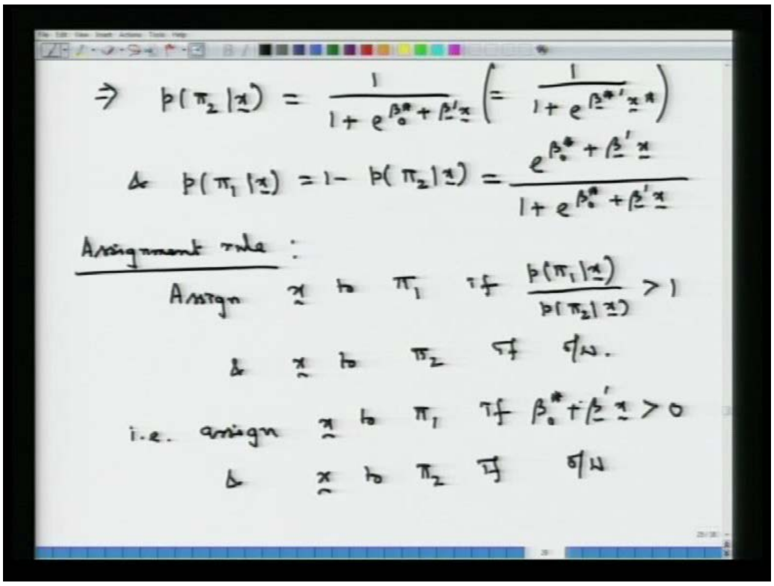
Now we had earlier denoted this p pi i to be equal to p i simply. So, for notational convenience, we are writing this as p i p of p pi i equal to p i. So, this log odds ratio, thus is going to be equal to these two terms cancel out, and what we will be having the log odds ratio as this is a p 2 times p pi 1 given x that divided by p 1 into p pi 2 given x.

And from the condition of the logistic discrimination, this is equal to our beta naught plus beta prime times x. Now, this further would imply that this ratio of this posterior probabilities p pi given x, this divided by p pi 2 given x quantity that is going to be equal to p 1 by p 2 times e to the power beta naught plus a beta prime x. Let us write this in the form that it is log of p 1 by p 2 term, this plus this beta naught plus a beta prime x quantity. And let us write this further as e to the power a constant beta star which is beta star is beta naught plus log of p 1 by p 2, this plus beta prime x wherein we have used the fact that is beta star, we are denoting this as beta naught plus log of p 1 by p 2.

So, this implies what we have here, now note that p pi 1 given x plus p pi 2 given x that is equal to 1. So, what we can write in place of p pi 1 given x is 1 minus p pi 2 given x that divided by p pi 2 given x. So, that is equal to e to the power beta star plus a beta prime times x. Now, either you write it in this particular form or one can write this as e to the power a vector beta star multiplied beta star prime that multiplied by x star, wherein what would be this beta star vector? The beta star vector would have a beta star in the first place and then that augmented by this old beta vector, and x star similarly is going to be equal to one augmented with this x vector, the original feature vector.

Because this one is going to get from this beta star and lead us to this beta star plus beta prime x. Now from this expression, what we have in here. It is further that what we have is one can express this ratio in terms of just p pi 2 given x.

(Refer Slide Time: 07:12)




So, this is going to imply that p pi 2 given x is going to be just equal to 1 upon 1 plus e to the power beta star, this plus a beta naught star, let me see the notation that I had used. Let us still write this as a beta naught star quantity, because in other two just have some understanding that this is what is corresponding to the constant term out here.

So, let us just keep this as a beta naught star. So, that we have this here. So, its beta naught star this plus beta prime x quantity either in this form or in the form of 1 plus e to the power beta star vector prime times x star, which are equivalent. So, when we have this posterior probability of the second population given x to be given by this, one can also see that what p pi 1 given x is. So, that is going to be 1 minus p pi 2 given x, and that from this expression is just going to be equal to e to the power beta naught star, this plus this beta prime x that divided by 1 plus e to the power beta naught star plus this beta prime x quantity.

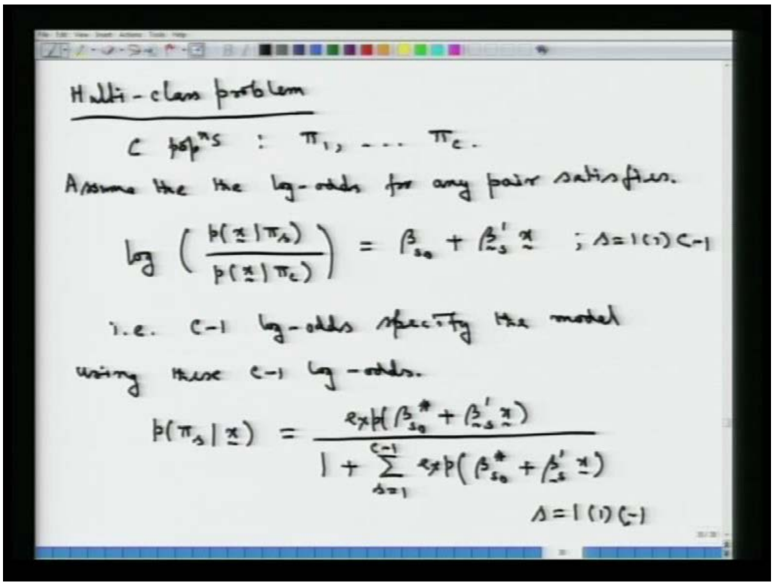
So, these are the two posterior probabilities of the respected population. So, we have obtained that it is p pi 1 given x is given by this and p pi 2 given x is given by this. So, the assignment rule or the discriminate discrimination rule is the following assign x to pi 1. If the posterior probability of pi population is higher than that of the second population, if that is assign x to pi 1 if we have got p pi 1 given x greater than p pi 2 given x or in terms of this odds ratio what we can say is that if this ratio is greater than 1 and x to pi 2 if otherwise.

So, that is basically the assignment rule that we get if we look at such a formulation that the log odds ratio satisfies that particular relationship between the log odds and that of the explanatory variables, the feature vector which is contained in x. Now a thing that should be noted here at this particular point of time, that when we are trying to implement this particular assignment rule in practice, this p pi 1 given x or p pi 2 given x depends on the parameters beta naught star. And the parameters which represent in this beta vector which is beta 1, beta2, beta p.

And from the given data, from the learning sample one has to thus at that particular stage come up with some estimates of these quantities. And based on the estimated these quantities will be actually implementing this classification rule. Now let us try to extend this approach, now it is another way to look at this particular term is one can see that, that is assign x to pi 1 if if we are looking at what is this ratio p pi 1 by p pi 2 given x, it is this divided by this particular quantity. So, it is going to be e to the power beta naught star plus beta prime x that is greater than 1. Or in other words we will have this rule to be given by this beta naught star plus this beta prime x, that to be greater than log of 1 0.

And x to pi 2, if we have the condition otherwise, that is in a more simple manner it is this straight forward depending on these unknown quantities p plus 1 of them a beta naught star and a and p quantities which are there in this beta vector.

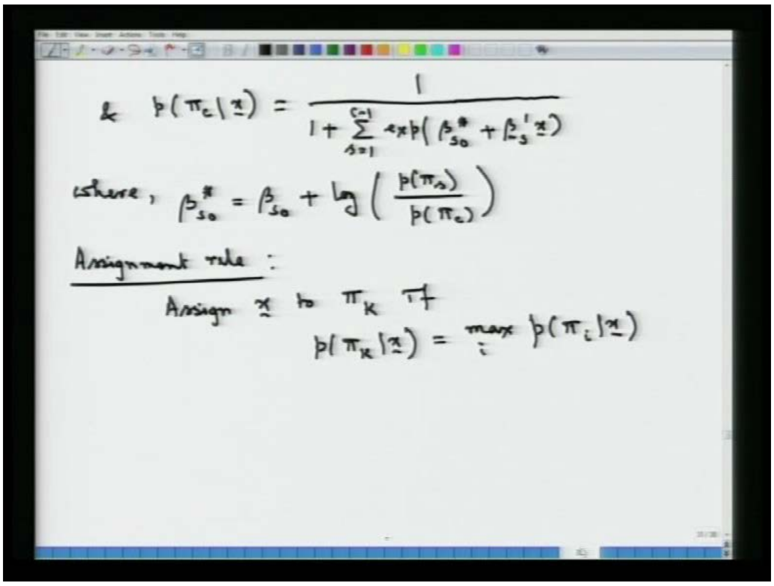
(Refer Slide Time: 11:48)



So, we now look forward to extending this particular logistic discrimination function in a c class problem a multiclass problem, wherein we assume that we have got c populations multiclass problem. And trying to frame, what is going to be the logistic discrimination rule in such a situation.

So, we assume that there are c populations, which are say denoted by pi 1 pi 2 pi c. Now what we are now going to assume is the following, assume that the log odds for every pair satisfies any pair of populations, satisfies the following relationship that the log of p pi p x given pi s say s th population, that divided by p x given pi c. This satisfies an equation say is equal to beta s naught plus a beta s vector transpose x, this is for s equal to 1 to up to c. So, we are looking at, this is the pair of pi c th population and any of the pi s population for i equal to 1 up to. I am sorry and for s equal to1 to up to c minus 1. So, we are looking at pairs of such populations, and then looking at the log odds of these quantities which we assume that it is of this particular form. So, this would imply that this c minus 1 log odds specify the model null. So, we have got this c minus 1 log odds, which we get from this expression for s equal to 1 to up to c minus 1, these c minus 1 log odds specify the model completely. Now, using these c minus 1 log odds and using the fact that summation of all these quantities p x given pi i summation from i equal to 1 to up to c, these are going to add up to one. What we can now see is the following, using this c minus 1 log odds, what will be getting is that p pi s the posterior probabilities p pi s given x, this is going to be given by e to the power beta s naught star, it is almost the same as what we had for the two class problem. That plus beta s transpose x that divided by 1 plus this summation of s equal to 1 to up to c minus 1 of these quantities e to the power beta s naught star, this plus beta s prime times x, and this is for s equal to 1 to up to c minus 1.

(Refer Slide Time: 15:26)

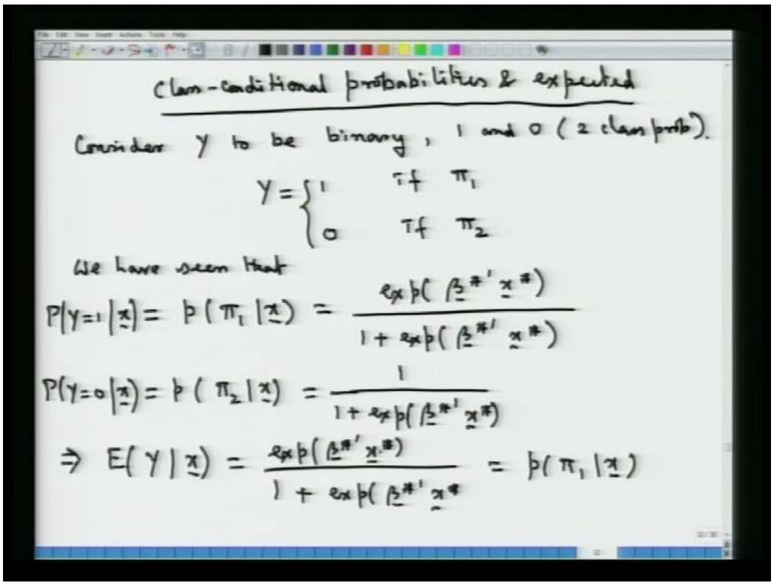



So, for these c minus 1 populations, the posterior density p pi s given x is going to be given by this. I will just define, what is beta s naught star and those quantities. And you will have this p pi c given x, that to be given by 1 upon the same denominator. So, this is 1 upon 1 plus summation s equal to 1 to up to c minus 1, this also was c minus 1 of e to the power the same quantity as before. So, that was denoted by beta s naught star, this plus this beta s prime x quantities. So, it is this term, wherein we have got as in the previous setup this beta s naught star, that is going to be given by beta s naught the first one, then that multiplied by log of p pi s which we have denoted by p s quantity simply that divided by p pi c.

Now, once we have these expressions, c of these out here. This is the posterior density of the s th population given x, this is for s equal to 1 to up to c minus 1, and this is for the c th population. Assignment rule is the following, using this logistic discrimination function is just the extension of what we have for the two class problem. Assign x to pi k if we have p pi k the posterior probability of the k th population given x, if that is maximum over i of all these posterior probabilities p I, given x right. So, this is what we have as the classification rule, we are going to assign x to pi k, if we have got this to be true.

Now, this once again can be expressed in terms of this beta s naught star and beta s vectors, because all these are that.

(Refer Slide Time: 17:51)




So, one can find out what is that quite easily. Now let us look at the following relationship, which is interesting the name why logistic discrimination comes at all, let us now look at the class conditional probabilities class conditional probability and expected response. So, what is the relationship between these, and how we are going to get the name logistic discrimination? Now for simplicity, consider a two class problem consider a Y a variable to be binary say taking value 0 and 1. So, this is corresponding to the two class problem.

Now, in what sense is this two class problem. So, Y takes the value say 1, if the population is pi 1 and takes the value 0, if the population is pi 2. So, corresponding to the particular feature vector, we of course have the in the pre classified examples in the learning sample, what is the class membership of that particular feature vector. And if that is pi 1, then the value of the Y variable which is a binary variable. We take that to be equal to 1, and if the membership of the feature vector is pi 2 population, then we take the value of this binary variable to be equal to 0.

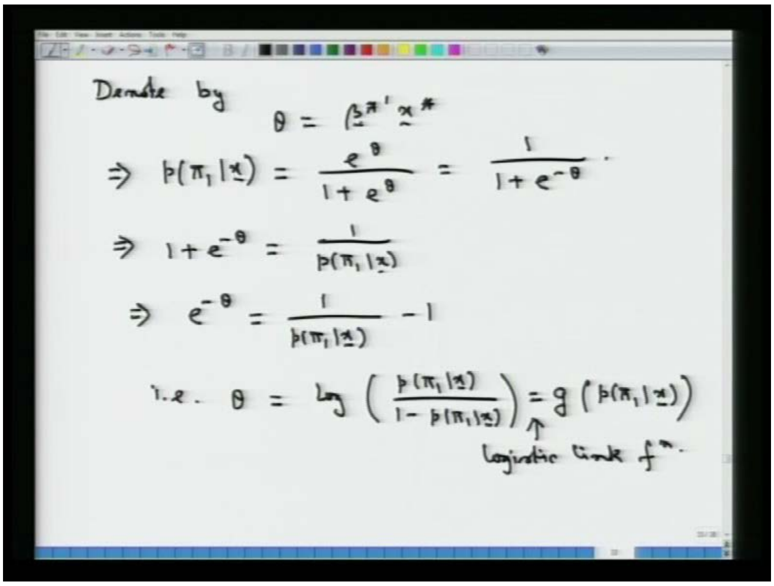
Now, what we have seen earlier is that, we have already seen that, now these are that is two class problems. Now we are going to write this expression of p pi 1 given x which we have discussed today, which is e to the power. Let us look back and see what it is for the two class problem, we had this p pi 1 p pi 1 given x to be given by this expression, and p pi 2 given x to be given by this expression. So, let us use those expressions and write this as a beta star prime x star vector that divided by 1 plus e to the power beta star prime x star vector.

And this quantity, if you now look at this binary random variable Y. So, this has got two values 1 and 0 corresponding to this. So, if we are looking at p pi 1 given x, then that is nothing, but in terms of the Y variable, its probability that Y is equal to 1 given x. So, this is what is a conditional probability of Y taking the value 1 given x is observed, and similarly we have also seen that p pi 2 given x that is equal to 1 upon the same denominator. So, that is 1 plus e to the power beta star prime this as x star.

Now, what is this equal to this in terms of the binary variable. This is probability that Y is equal to 0 given x. So, we if we have got these two as the conditional probability masses of Y taking the value 1, and Y taking the value 0. This would imply that if we look at the conditional expectation conditional expectation of Y given x, what is that going to be equal to? This takes the value 1 with this probability and 0 with this probability. And hence, the conditional expectation of Y given x is just going to be given by this particular expression, which is e to the power beta star prime x star that divided by 1 plus e to the power beta star prime x star, which by the way is nothing, but p pi 1 given x.

Now, from this expression here, which we have got this conditional expectation to be equal to this term here. Let us now denoted this quantity beta star prime x star to be equal to a quantity which is theta, denote by theta the quantity which is beta star prime x star.

(Refer Slide Time: 22:04)

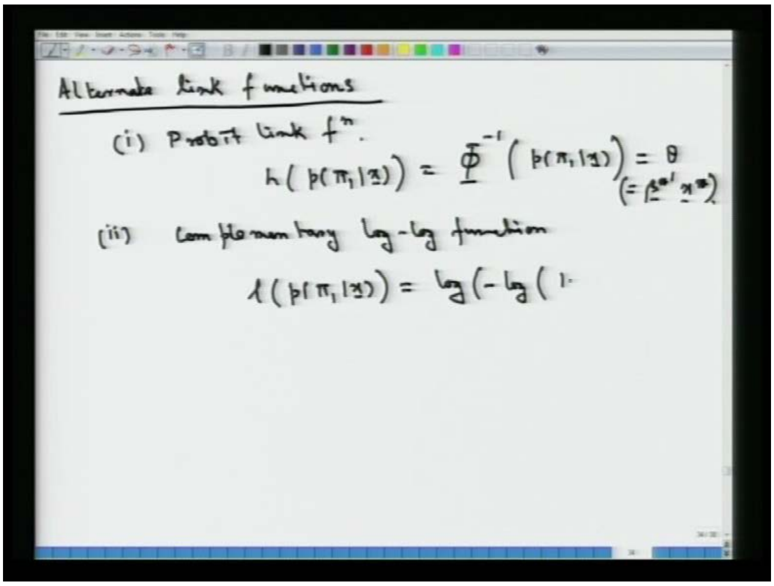



So, this would imply that what we have as p pi 1 given x is nothing, but equal to, let us see this expression it is going to be equal to e to the power theta divided by 1 plus e to the power theta. So, that is equal to e to the power theta that divided by 1 plus e to the power theta, which one can also write as one upon 1 plus e to the power minus theta.

So, if we have this expression, this would imply that 1 plus e to the power minus theta that is equal to one upon p pi 1 given x. So, this would imply that e to the power minus theta is equal to one upon p pi 1 given x, this minus 1. That is, one can write the theta quantity in terms of log of this other way round. So, it is going to be equal to log of p pi 1 given x this divided by 1 minus p pi 1 given x, now this is some function of p pi 1 given x.

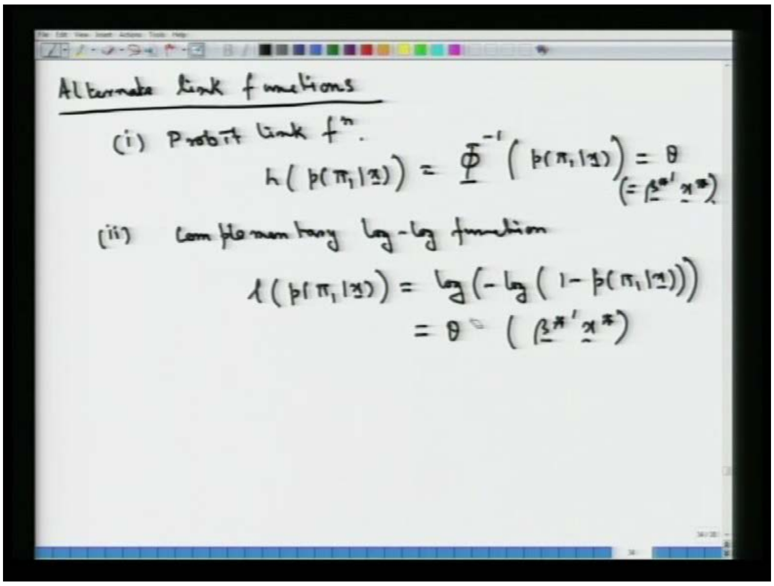
Now, since we have got this function form in terms of this log odds ratio here, and that p pi 1 given x is of this logistic function. We have the name that, this link that we are linking theta with p pi 1 given x the posterior density of pi 1 given x through a logistic link function. So, this is the logistic link function, and hence what we get the name as the logistic discrimination. Now alternate forms of having this link function, because this is just the relationship between this theta, which is beta prime beta star prime x star and this p pi 1 given x.

(Refer Slide Time: 24:36)




So, if this function is of the form that it is log of p pi 1 given x or rather it is assumed to be of the form that it is log of p pi 1 given x, that divided by 1 minus p pi 1 given x. Then what we get is the logistic discrimination alternate link functions are following.

(Refer Slide Time: 24:35)

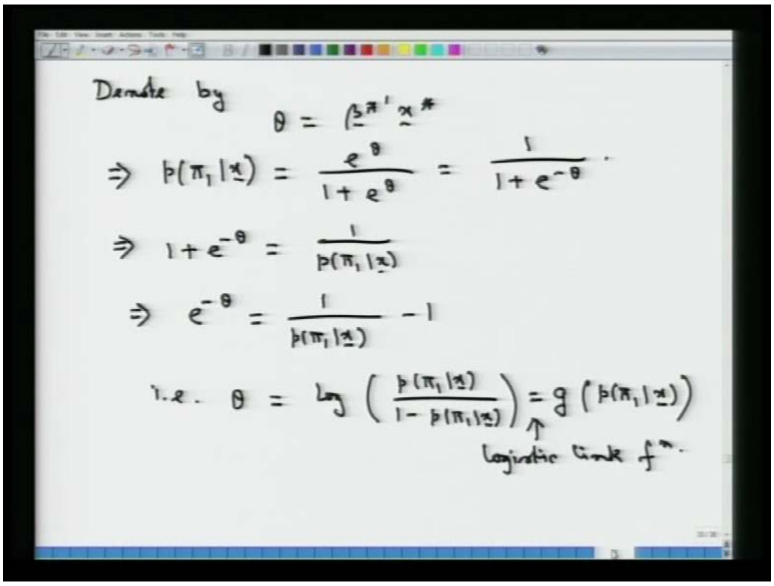



Alternate link functions, there are two popular alternate link functions. One which is called the profit link, profit link function wherein we assume that this function, what we have a function of p pi 1 given x this is a probability. So, this h of this p pi 1 x is going to be given by capital phi inverse of p pi 1 given x.

So, this particular function here phi inverse, where capital phi is the probability distribution function of a standard normal vitiate that is equal to theta, which is our x star beta star prime x star. So, this is equal to our beta star prime times x star. Now the second type of popular link function is what is called the complimentary log function, a complimentary log function log log function rather, wherein we assume that we have got a function of this p pi 1 given x to be given by the following quantity which is log of minus log of 1 minus p pi 1 given x.

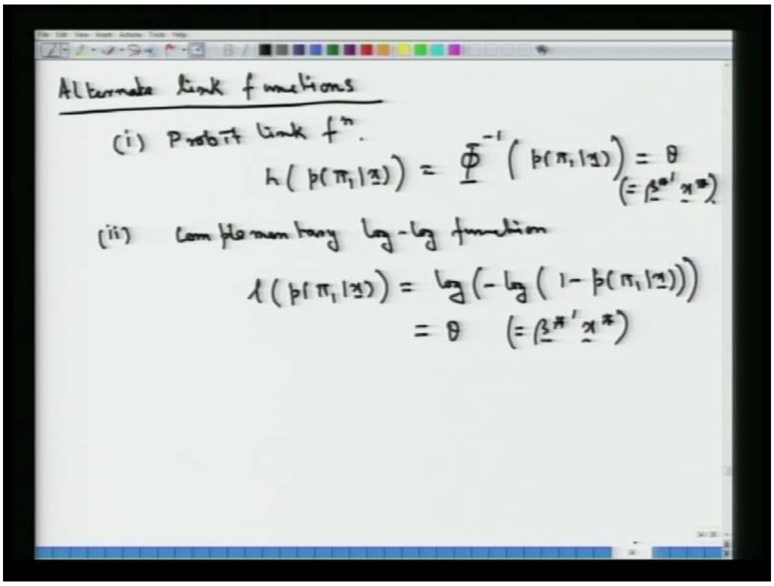
So, that this bracket ends here and that is linked with this theta, which once again is that linear combination of the parameters with the feature vector with the constant vector 1 attached to it.

(Refer Slide Time: 26:28)



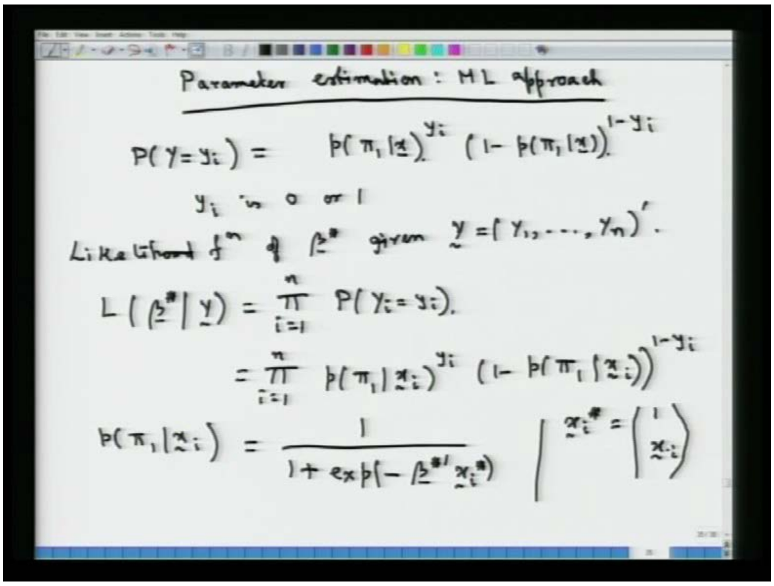

So, that these are two alternate link functions. In case of a logistic discrimination, the link that we have is precisely this that the functional link between p pi 1 given x. And theta is given by this expression which is logistic link, this is the profit link. And this is the complimentary log log function link.

(Refer Slide Time: 26:40)




Now let us last look at or rather talk about little bit about parameter estimation.

(Refer Slide Time: 26:52)




Now the method of parameter estimation that is usually adapted in such situation is the method of maximum likelihood. So, parameter estimation M L approach let us still look at that simple formulation that we have got two possibilities 0 and 1. So, we have got under such a situation of the previous setup, what we have discussed here in this formulation that we have got Y i es to be defined in this particular way, and in such a situation what we are going to have is the following.

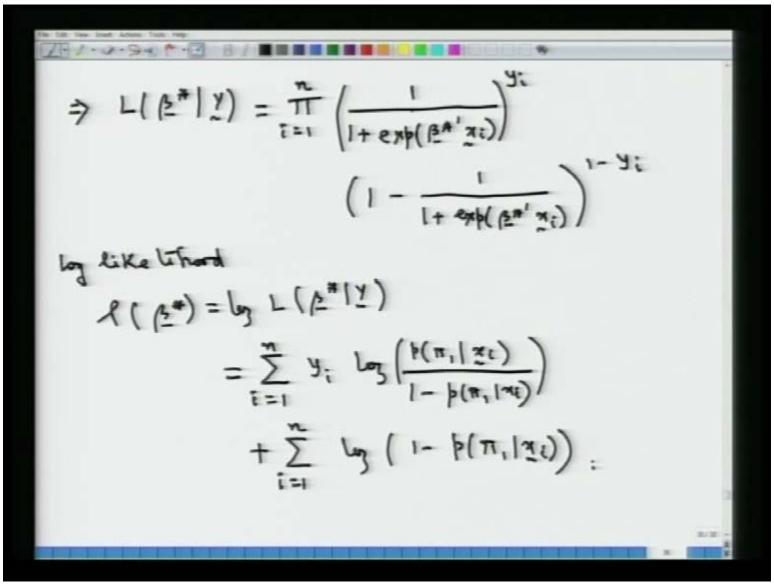
Probability that Y is equal to y i that is going to be given by p pi 1 given x, this to the power y i into 1 minus p pi 1 given x, this to the power 1 minus y i. Now, y i takes either of the two values y i is 0 or 1. That is we are looking at this of course, is the conditional quantity conditional probability mass function given x. So, this is y equal to 0 is what we will have as 1 minus p pi 1 given x that is p pi 2 given x. And probability that y is equal to 1 given x that is equal to p pi 1 given x. So, this is that particular quantity.

Now, hence if we have y 1, y 2, y n the likelihood function, the likelihood function of the parameters, which basically is coming in that beta star vector, given this y vector equal to y 1, y 2, y n, this is for n random samples. We are going to have this as say we denote this by l beta star vector, this given y vector. This is going to be equal to, because all the y is that is what we have they are independent. So, this is going to be the product of i equal to 1 to up to n, probability that y i is equal to small y i.

So, what y i is basically denoting the ith record in the data, and that in the ith record in the data, we will have the corresponding feature vector to be denoted by x i.. So, this is going to look like the following, that it is product of i equal to 1 to up to n the product of these quantities keeping in mind that, when we are looking at probability that y i is equal capital y i is equal to small y i, this is pi 1 given x i vector that to the power y i into 1 minus p pi i i am sorry this is not pi i this is pi 1 one minus p pi 1 given x i that to the power 1 minus y i.

Now, we use the fact that this this p pi 1 given x i is nothing, but this is going to be of the form that it is in the form 1 plus e to the power minus beta star transpose into x i star quantity, wherein x i star is the following term x i star vector, is the vector which is one in the first entry. And then we have x i the feature vector to make up the rest of the p entries in here. So, using this particular fact, we have this and accordingly if we plug-in the values of p pi 1 given x i in this expression here. We have the explicit form of this likelihood.

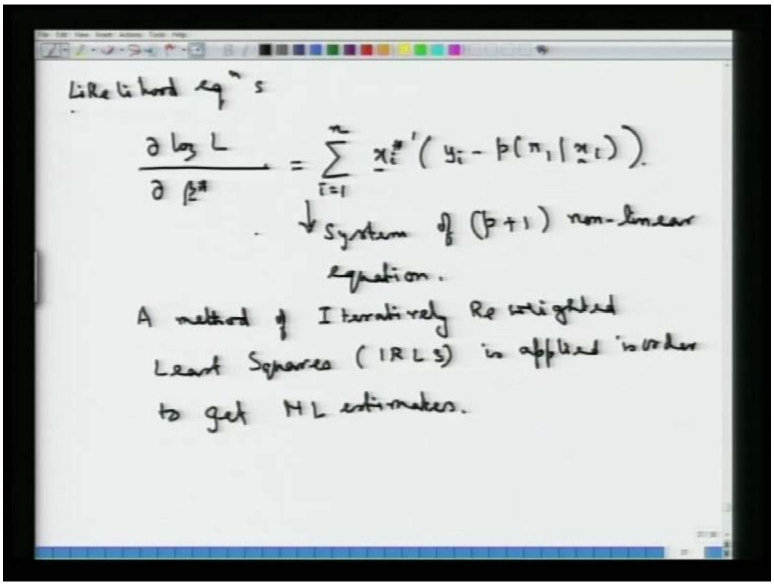
(Refer Slide Time: 31:06)




So, this will imply that this l beta star given this y vector that is going to be equal to product i equal to 1 to up to n, and then we have p pi 1 given x. So, that now we are writing that as one upon 1 plus e to the power beta star prime x i quantity. So, this is p pi 1 given x that to the power y i that multiplied by 1 minus this quantity. So, it is 1 minus one upon 1 plus e to the power the same quantity. So, just erase this 1 here. So, that we have a big denominator coming up. So, its beta star prime x i vector that to the power 1 minus y i.

Now, this if this is the likelihood, one can also write the log likelihood. The log likelihood function, let us denote that by small l beta star vector, which is log of this l beta star given y expression. And that can be written in terms of this compact summation, which I will just write in here which is summation i equal to 1 to n y i times log of in term I am just keeping it in terms of p pi i. So, that the expression is not too messy, this is going to be given by the following quantity which is this log of p pi 1 given x I, this is combining the second term also. So, this term plus summation i equal to 1 to up to n of log of 1 minus p pi 1 given x i term.

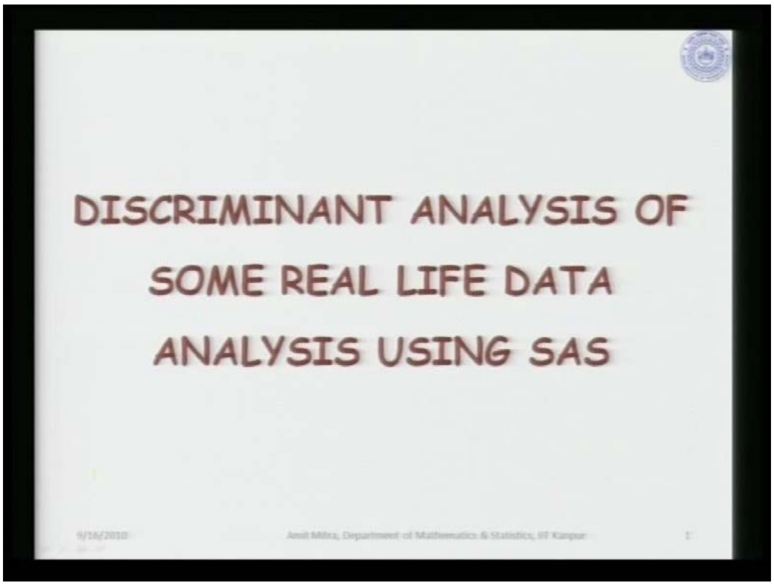
(Refer Slide Time: 33:24)




So, this is the log likelihood, then we can get to the likelihood equations from here likelihood equations corresponding to the p plus 1 parameters. We can write that compactly in the following form that this del log l with respect to this beta star vector is going to be equal to x i star transpose that into y i minus p pi 1 given x i expression. Now, this is going to be a system of non-linear equations. So, this is what we are going this summation is over i equal to one to up to n. So, this is a system of p plus 1 non-linear equations, and of course no closed form solution exists for such a system of non-linear equations. A method which is called an iteratively reweighted least squares is applied, a method of iteratively reweighted least squares or I R L S is usually applied in order to get the maximum likelihood estimates.

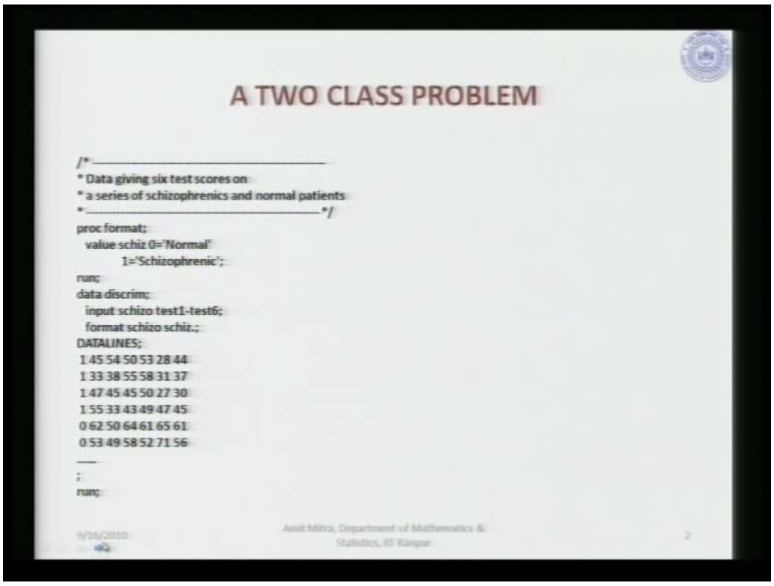
So, technically using this system of p plus 1 non-linear equations, and using I R LS, one gets to the maximum likelihood estimate, estimates of the p plus one unknown quantities. Now, once you have p plus 1, unknown quantities estimated, then one can actually look at implementation of this entire logistic discriminate function. Now, what we are next going to do is we are going to look at some real life data, and we are going to apply the type of discrimination analysis methods that we have learnt in the theoretical classes, in order to see what sort of discrimination we get in practice.

(Refer Slide Time: 35:50)




So, we now look at this a small presentation which is on a power point.

(Refer Slide Time: 36:07)




So, we are going to now look at, some real life data and real life data analysis wherein we are going to implement the type of methods that we have learnt in discriminate analysis. Now, the data analysis is done using a SAS routine. Let us look at the first example, in the first example we have a two class problem. So, there are two populations, now the two populations are following that it is a set of patients. The first populations of patients are normal patients and the other type of population is the type of patients which are schizophrenic.

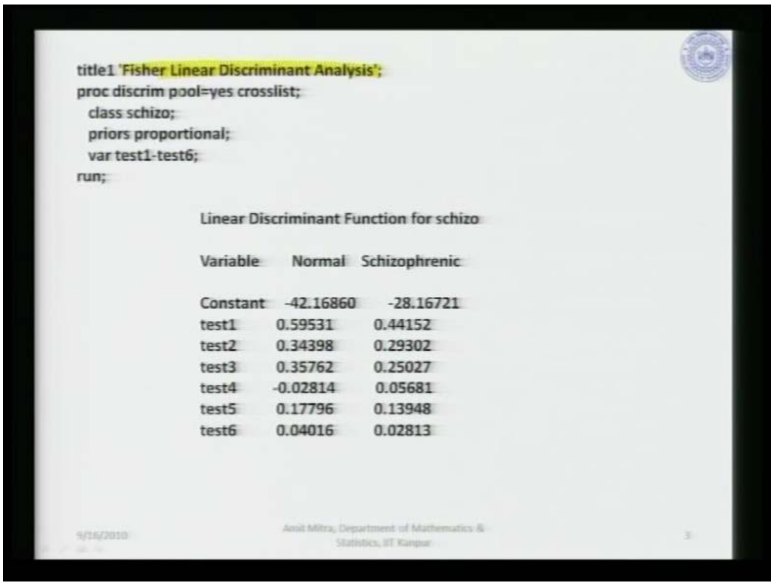
So, we have got this these two classes, now we have the data, the data is giving six test scores on the series of schizophrenic and normal patients. Now, this is the format of the data, it is a huge data. So, dots are given. So, it is basically this type of data. So, it represents what the data represents is that the first entry in each of the rows this quantity here, it is basically the class identification. So, this is the data which is the which is comprising of the learning set data. And it has got class identification tags and hence this row of the records is what is corresponding to a schizophrenic pair patient.

Now, we are denoting that 1 if the patient is schizophrenic, and 0 if he is normal. So, we have got this to be the class membership and next six entries this 45, 54, 50, 53, 28 and 44 are the test scores corresponding to that particular patient. So, we have records like that in the data. So, some of them has class identification 1, some of them has class identification 0. So, these are normal individuals, and these are schizophrenic individuals, and these are the corresponding test scores.

Now, this six dimensional vector of test scores is now going to correspond to what we have as a feature vector. Now given this liming sample to us, we would construct discriminate functions and classification rules based on this discriminant functions, such that we will once again look back at the learning sample. And then see how that constructive discriminant function is able to classify the pre classified examples, how it how it is performing on the data that is what we have in the learning sample. And then, that would lead us to the desired classification functions.

Now, the in the first case, we apply a type of discriminant function that we had studied in the very first lecture in discriminant analysis, which is a fisher linear discriminate function.

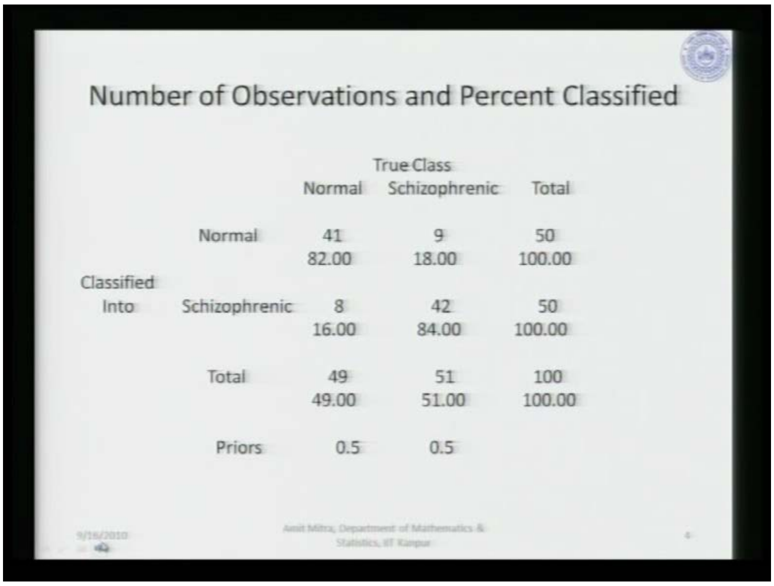
(Refer Slide Time: 38:45)




So, it is implemented using a SAS code, which uses the procedure disc rim, and then looks at this particular data. Once we apply the fisher linear discriminate function to the data that we have, we get the linear discriminate function scores here, which are the coefficients that we are going to get, because this is what we have the constant term, and these are the coefficients corresponding to each of the feature vectors.

Feature vectors components are test 1, test 2, test 3, test 4, 5 and 6. And this is for the schizophrenic patients.

(Refer Slide Time: 39:18)



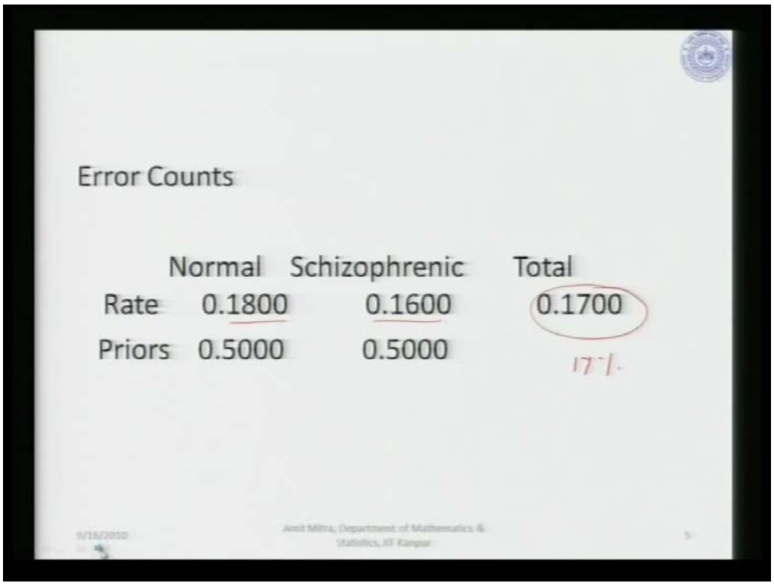
So, these are the two different functions that we are going to get the function coefficients for the two types of patients. Now, using the fisher linear discriminate function, we look at what type of or what is the number of observations. In the learning sample, how is the performance on that learning sample of the constructed discriminant function. Now, what we get here is that now this table here is what is the confusion matrix and along with the confusion matrix, this also gives us the percent of observations which are correctly or wrongly classified.

Now, on this side we have got here. So, we have got the true class membership. So, the true class is either its normal or it is schizophrenic, and that true class observation where it is testing classified into. So, an observation coming from normal class can get classified either to the normal class or it can get classified to the schizophrenic class in which case it is going to be a misclassification. So, what we have from the given data and fisher linear discriminate function is that 41 cases, which had a true class membership of normal are now classified as normal. What is the total number of such normal patients in a class of now, total data sizes 100, among that 49 are normal patients normal individuals rather and 51 individuals are are are schizophrenic.

So, from among 49 normal individuals true class membership number 49, we have 41 of those been classified in to the normal category. So, this is a correct classification of the normal category individuals. Among those 49 individuals, 8 individuals among the 49 normal individuals 8 have been misclassified in to coming from the schizophrenic class. So, these are misclassifications. Now, if we look at the other class, true class membership is schizophrenic. There are 51 such patients, now from among those 51 patients; we are classifying 42 of them in to the class which is schizophrenic. And hence we are what we are doing here is a correct classification. And from among this 51 schizophrenic patients, 9 of them are classified as coming from a normal category of patients, normal category of individuals rather and hence this is misclassification.

So, this is what is giving us the confusion matrix, after we apply the fisher linear discriminant function. Herein, these two are the correct classifications, and these two are the wrong classifications.

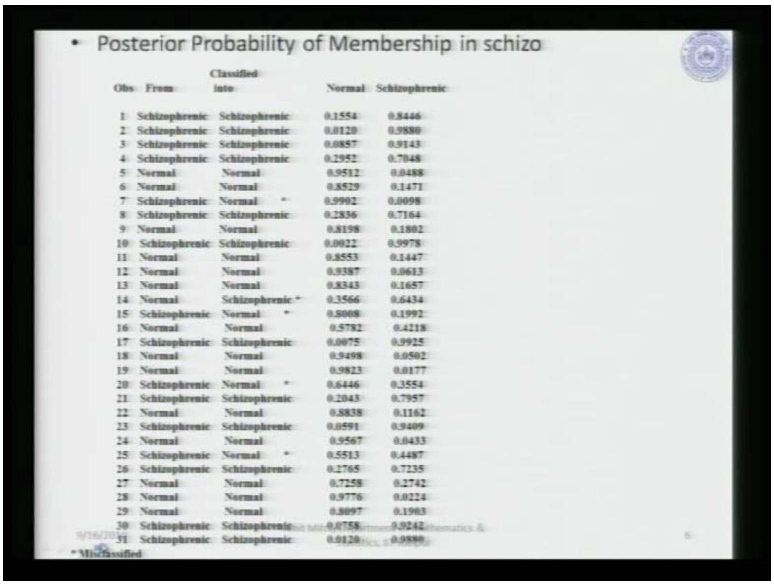
(Refer Slide Time: 42:50)




So, we have got 18 percent of a total of 50 classifications made in the wrong category and 16 percent here in this row here going to the wrong category. And hence, now we assume here equal priors that it is 0.5, 0.5 for the two populations. Now the error counts, thus from the 2 classes pulled up its 18 percent from the normal class, 16 percent from the schizophrenic class. And hence, it is basically this percent of the observations that is 17 percent of the observations are wrongly classified using this classification rule. That is, it is simple to see that out of 100, these 17 cases are misclassified.

So, that is what is corresponding to a fisher linear discriminant function. Now corresponding to the fisher linear discriminant function, if one looks at the posterior probability of membership in to the class schizophrenic.

(Refer Slide Time: 42:52)



- Posterior Probability of Membership in schizo

| Obs | From | Classified into | | Normal | Schizophrenic |
|---|---|---|---|---|---|
| 1 | Schizophrenic | Schizophrenic | | 0.1554 | 0.8446 |
| 2 | Schizophrenic | Schizophrenic | | 0.0120 | 0.9880 |
| 3 | Schizophrenic | Schizophrenic | | 0.0857 | 0.9143 |
| 4 | Schizophrenic | Schizophrenic | | 0.2952 | 0.7048 |
| 5 | Normal | Normal | | 0.9512 | 0.0488 |
| 6 | Normal | Normal | | 0.8529 | 0.1471 |
| 7 | Schizophrenic | Normal | * | 0.9902 | 0.0098 |
| 8 | Schizophrenic | Schizophrenic | | 0.2836 | 0.7164 |
| 9 | Normal | Normal | | 0.8198 | 0.1802 |
| 10 | Schizophrenic | Schizophrenic | | 0.0022 | 0.9978 |
| 11 | Normal | Normal | | 0.8553 | 0.1447 |
| 12 | Normal | Normal | | 0.9387 | 0.0613 |
| 13 | Normal | Normal | | 0.8343 | 0.1657 |
| 14 | Normal | Schizophrenic | * | 0.3566 | 0.6434 |
| 15 | Schizophrenic | Normal | * | 0.8008 | 0.1992 |
| 16 | Normal | Normal | | 0.5782 | 0.4218 |
| 17 | Schizophrenic | Schizophrenic | | 0.0075 | 0.9925 |
| 18 | Normal | Normal | | 0.9498 | 0.0502 |
| 19 | Normal | Normal | | 0.9823 | 0.0177 |
| 20 | Schizophrenic | Normal | * | 0.6446 | 0.3554 |
| 21 | Schizophrenic | Schizophrenic | | 0.2043 | 0.7957 |
| 22 | Normal | Normal | | 0.8838 | 0.1162 |
| 23 | Schizophrenic | Schizophrenic | | 0.0591 | 0.9409 |
| 24 | Normal | Normal | | 0.9567 | 0.0433 |
| 25 | Schizophrenic | Normal | * | 0.5513 | 0.4487 |
| 26 | Schizophrenic | Schizophrenic | | 0.2765 | 0.7235 |
| 27 | Normal | Normal | | 0.7258 | 0.2742 |
| 28 | Normal | Normal | | 0.9776 | 0.0224 |
| 29 | Normal | Normal | | 0.8097 | 0.1903 |
| 30 | Schizophrenic | Schizophrenic | | 0.0758 | 0.9242 |
| 31 | Schizophrenic | Schizophrenic | | 0.0120 | 0.9880 |

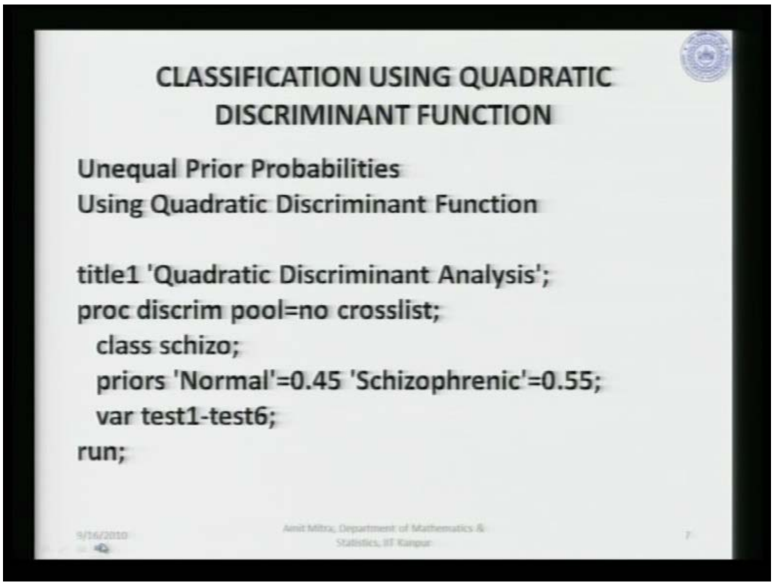
* Misclassified

So, schizophrenic was a class which was having the class membership as 1. So, we have got these are the observations, these are the observation numbers. And the individual 1,the observation number 1 is coming from the class schizophrenic, it has got the posterior probabilities of the of in schizophrenic population as 0.84 and the posterior probability of a normal population, normal individual population is 0.15.

So, we see that this posterior probability is higher and hence what we have this observation classified correctly in to the schizophrenic class here. So, same as the interpretation for each records, each row of the records here say for example, if you look at the fifth record here, the fifth case is coming from a normal category of individuals. Now the posterior probabilities are coming out as 0.95 for the normal class and 0.04 for the schizophrenic class and since this is higher we classify correctly in to the normal class.

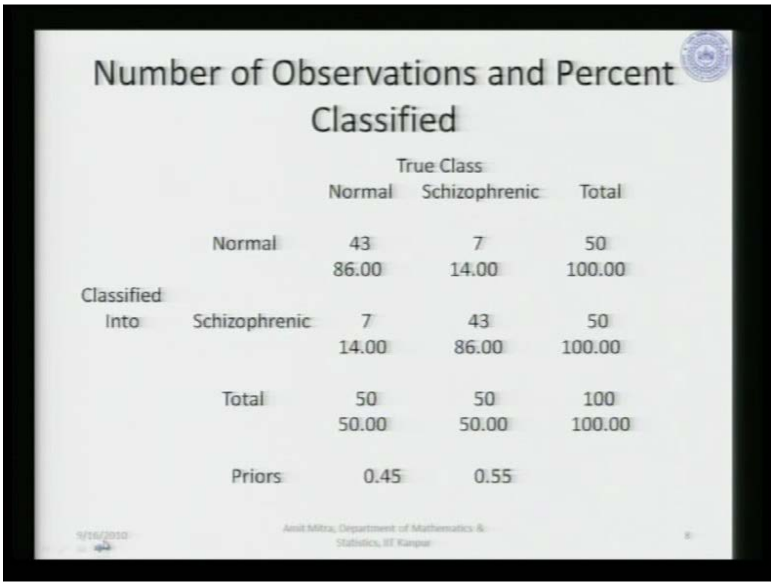
So, all these are correct classifications up to this particular point here. Now in the case number 7, there is a schizophrenic patient it is what the class membership is schizophrenic. And we are the posterior probability of the normal individual category as 0.99, and in the schizophrenic category the posterior probability 0.0098. And since, this is higher, we classify it as coming from a normal category of patients. However, that is a misclassification. So, this classification is wrongly done out here. So, these are the cases wherein we have got the misclassifications. So, all these stars here indicate that the observations are going to misclassify based such posterior probability, the computed posterior probabilities.

(Refer Slide Time: 44:56)



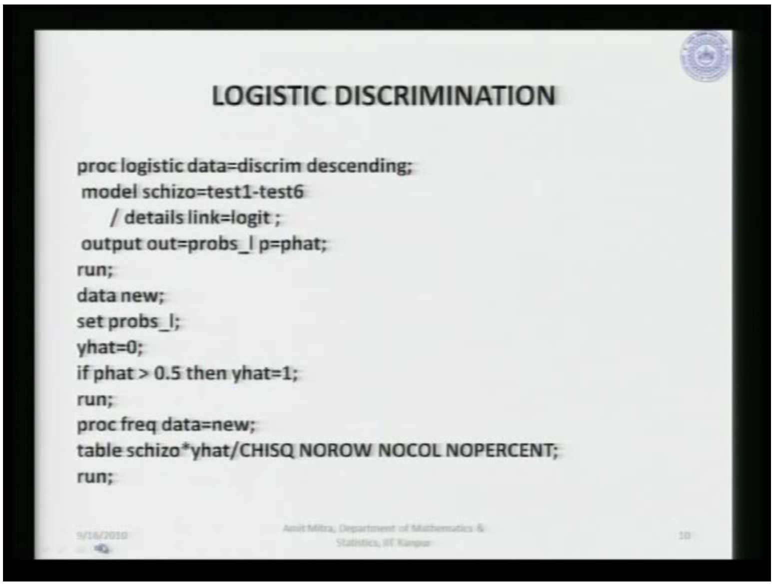

Next we look at the same data set only, and use a quadratic discriminant function with an unequal prior probabilities. So, this is what we are now looking at with unequal prior probabilities, and we are looking at a quadratic discriminant function. We take the 2 priors 0.45 and point 4 or 0.55.

(Refer Slide Time: 45:24)



So, these are the 2 prior probabilities. And with a quadratic discriminate function, we have the following confusion matrix along with the percent classifieds. So, this is once again, this is a true class membership which are 2 types normal, schizophrenic. And an individual coming from whichever class has got the possibility that it gets classified either in to normal or it is schizophrenic class.

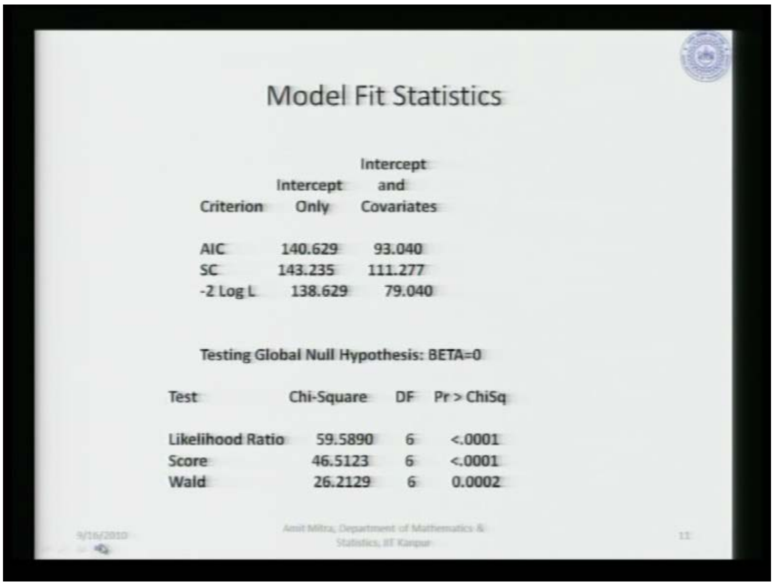
Now, we see that from among 50 observations, now 43 in the normal category are correctly classified, 43 of them are also correctly classified. Then we have the number of observations coming here 50 from this class. And we have these numbers which are this classification numbers. So, these are the 2 quantities, wherein we have got these classifications. So, the percent of this classification here as we see its 14 percent out of this total number of cases. The similar is the interpretation, when we look at once again the posterior probability membership of membership in to the class schizophrenic.




Now, we have once again based on such posterior probabilities the classifications, we have misclassifications in some cases, total 17 in all. So, these are misclassification (( )) the others are getting correctly classified using this quadratic discriminant function. We also apply a logistic discriminant function, and we look at what does the logistic discrimination that we learnt in today's lecture is going to lead us to. Now we use the proc logistic of the SAS procedures, in order to give this particular exercise with a link function as a logit link function, the type of link function that we have just now discussed, it is going to be that log of p pi 1 given x divided by 1 minus p pi 1 given x.

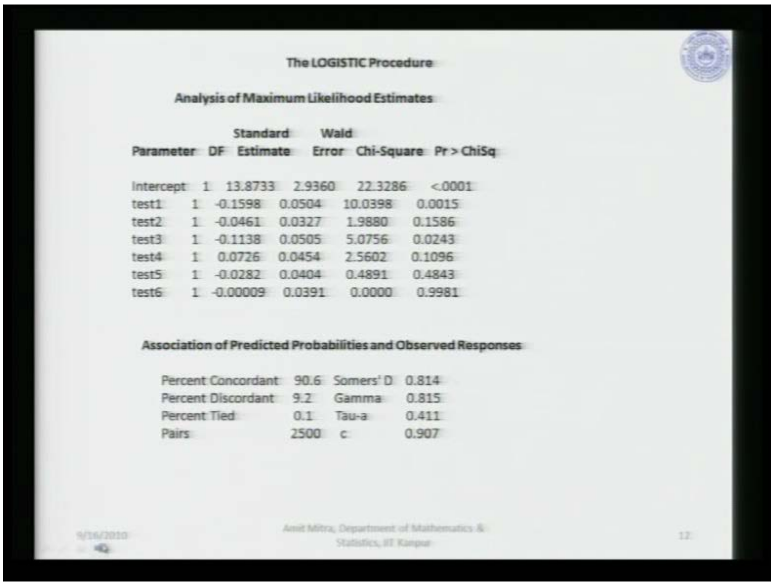
So, that is the with a logit link function, we are going to have these being classified. We take a cut off probability for a classification as 0.5. So, if the predictive probabilities are greater than 0.5, we take y hat the predicted class membership to be equal to 1. And if it is otherwise, we take the predicted class membership to be equal to 0. This is associated with the schizophrenic class, and this is associated with the normal class. Now further more, once we have the classification done up to this particular point then the classification is done, we will look at the confusion matrix. And then, we will look at applying the a frequency procedure in order to look at the measures of association between such predicted memberships and the actual memberships.

(Refer Slide Time: 48:05)




So, these are elementary model fit statistic, this is the a L C S C Schwarz criterion minus two log l criterion. So, the log likelihood type of framework, now these are the hypothesis testing for the logistic regression setup, wherein we will we are testing beta equal to 0. That is the hypothesis of interest these are the various tests likelihood ratio test, the score tes,t and the welds test. Each of them giving us a probability, which is very small less than 0.5.

(Refer Slide Time: 48:39)



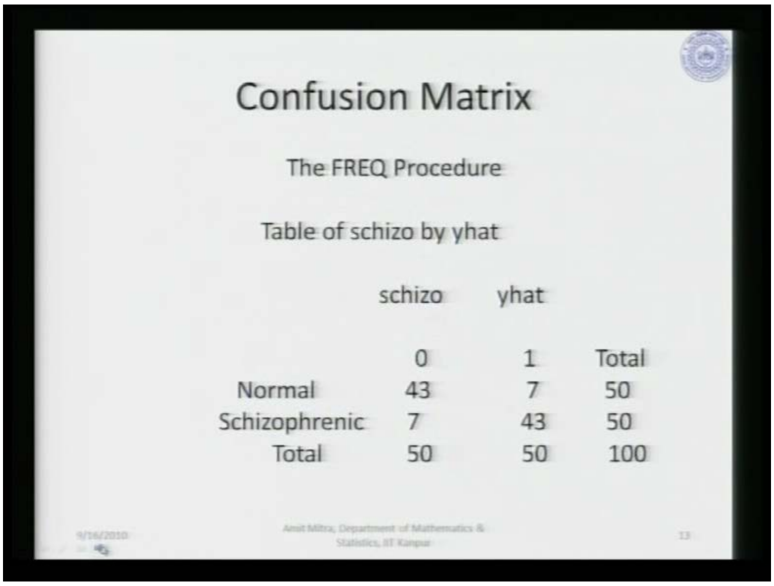

And what we have is these are the coefficients corresponding to such logistic regression of each of these tests, we assume that there is a constant term present.

So, these are the parameter, where the parameters are estimates this just shift a little bit, this is not standard estimate. This is to be taken with this so, it is standard error and welds chi square. So, this has to be coupled with this one. So, these are the parameter estimates, what we get from the data using an I R L S. And these are the corresponding weld the standard error quantities. So, these are the standard error columns and the welds chi square are these quantities. For those, which are less than a particular desired level of significance say 0.5, we reject the null hypothesis for all those terms there. So, the null hypothesis that this is equal to 0 is rejected, this is rejected this at 2 percent, so this is rejected. All other hypothesis are accepted at a 5 percent level of confidence.

Now, we also look at the association of the predicted probabilities through what when we are looking at the predicted class memberships as in after the predicted probabilities, if the predicted probability is greater than 0.5, we classify it in to say having the value equal to 1, y equal to 1 and 0 if it is otherwise. And then once again, we will be getting a confusion matrix. And from there, we look at the association of the predicted probabilities and the observed responses these are some standard measures of association, this is the percent concordant data. Ah after we have done the classification, this is a percent discordant in the data.

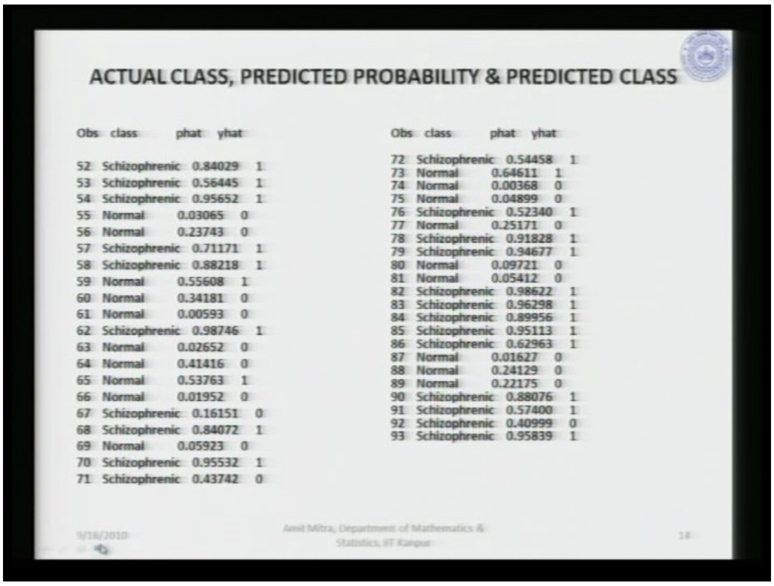
This is just an tied is 0.1 percent. So, we have a good fit, actually giving us percent concordant to be 90.6. So, these add up to 99.9, there is some round off somewhere. So, that 100 percent is not coming from all these cases. There are these many pairs, the Some are the criterion for measure of association is high, the gamma coefficient is high, tau a coefficient is high, the c kappa coefficient is also pretty high. So, the two the predicted probabilities and the observed responses, we naturally require them to be highly associated in order to have the classification to be worthwhile.

(Refer Slide Time: 51:00)




And we have that here, now this is the confusion matrix what we have in here, it gives us once again 43 correct classifications from the normal category, and 43 correct classifications from the schizophrenic category. And these are the observations which are wrongly classified coming from the two different classes. This is the y hat quantity and this is the actual quantities, actual class memberships.

(Refer Slide Time: 51:27)




So, what we have here is just a representative of the data after the model has been fitted. So, this is an observation number 52, it is a schizophrenic class membership. We have the predicted probability. Now, what is this predicted probability, this is probability that y is equal to 1, that is it is schizophrenic given x. So, that probability is 0.8402. So, it is greater than 0.5 and hence the predicted class membership is given as 1.

Now, this is a correct classification. Similarly, we have all these predicted probabilities, which is probability of y equal to 1 y i rather, corresponding to this case y i equal to1 given x quantities, and accordingly we have these. In situations, where this predicted probability is less than or equal to 0.5 as in this particular case, we will have that being classified in to the y hat category as taking the value 0.

So, that is in the normal category of patients. So, this is correctly classified, this is correctly classified. This also are correct classifications, correct classifications; however, this is a wrong classification, because we have this probability greater than 0.5. We classified that as y equal to 1 that is a predicted class membership is schizophrenic, which is wrong because the actual membership is normal. And hence, we have all other records in a similar way. Now for a same data set, we also apply a nearest neighbor classifier that we have learnt.

So, once again a proc discriminant, disc rim is used from SAS procedures with this non parametric method. Because it is a non parametric method, it is a nearest neighbor classifier; we use the same data and get results. So, these are the nearest neighbor classifier examples with Euclidean distances, these are the cross validation results. It is interesting to look at what is the confusion matrix, and how it is behaving. So, we see that, this is what we have the true class membership 41 out of 49 are correctly classified, 44 out of 51 are correctly classified, and these are wrong classifications using a nearest neighbor classifier. Now, we have a multiclass problem, maybe we will take it in the next lecture.